DOCUMENT RESUME

ED 287 287 FL 016 943

AUTHOR Ross, Steven
TITLE An Experiment with a Narrative Discourse Test.
PUB DATE 87
NOTE 11p.; In: Language Testing Research; Selected Papers from the Colloquium (Monterey, California, February 27-28, 1986); see FL 016 938.
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Audiolingual Skills; College Freshmen; *English (Second Language); Foreign Countries; Higher Education; Japanese; *Language Proficiency; *Language Tests; *Narration; Rating Scales; Screening Tests; Second Language Instruction; *Test Format; Testing Problems; Test Interpretation; Test Reliability
IDENTIFIERS *Japan

ABSTRACT

A study investigated the use of a narrative discourse task to test oral English proficiency in non-native speakers for screening and placement. The subjects were Japanese university freshmen entering a five-level course in spoken English. The subjects were shown an animated cartoon of a Japanese folk tale, with narration in Japanese, and later asked to narrate the story in English. The recorded speech samples were rated for pronunciation, accuracy, and fluency, and T-unit analyses were performed on a portion of the recordings. The results of the narrative task analysis were compared with the results of written tests. The findings suggest that this form of recorded narrative discourse can be useful in screening and placement, and further research on the length of speech samples necessary for adequate analysis is planned. Care must be taken to provide all students with appropriate background information for the narrative task. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED287287

AN EXPERIMENT WITH A NARRATIVE DISCOURSE TEST

Steven Ross
Osaka, Japan

With approximately forty-three percent of Japan's high school graduates going on to higher education, the study of foreign languages in Japan has reached a level of development comparable to the country's post-war economic growth. An important reason for the steady growth in foreign language education has been the college entrance examination. English as a foreign language is one of the essential parts of the national university standard examination, and is often the most important aspect of entrance examinations in the humanities departments of private universities and liberal arts colleges.

While the emphasis on English as a foreign language has grown, the basic methods of testing have not changed. Grammatical analysis and linear translation are still the preferred methods of evaluation, with little emphasis on listening comprehension and none at all given to speaking skills. These methods of testing have had a large impact on curriculum design in Japan, and the average college freshman is likely to have studied English as a foreign language for at least six years without ever having an introductory course in spoken English.

By the time the high school graduate has successfully entered a university, his or her first oral English course may be offered. A typical introductory course might be taught by a native speaker to a class of thirty to forty students. The disproportionate number of students to teacher in these college level classes precludes any serious attempt to test speaking proficiency directly. Often there are no provisions made to assess speaking proficiency at all at the university level unless an individual can pass qualifying examinations--again examinations that measure analytic ability and not oral skills.

Such examinations are commonly set by publishing houses or professional testing services. By far the most popular college level test, produced by the Society for Testing English Proficiency (STEP), does not directly measure speaking ability in the one-to-one interview format common in Britain and North America. Because of the disproportionate number of applicants to native interviewers, groups of students are typically confronted and asked to speak in turn on topics chosen by the interviewer, or on topics chosen at random from a pool of items. This approach results in a highly selective pre-oral examination screening and a method of speech elicitation that does not provide much more of a profile of an individual's sociolinguistic or discourse competence than would a conventional paper-and-pencil test of the individual's linguistic competence. Given this state of affairs, a practical method of oral language proficiency measurement which can reliably assess a foreign language speaker's ability to narrate under a reasonable time constraint can potentially serve to provide a more direct appraisal of the speaker's linguistic proficiency. Such a test could replace the paper-and-pencil screening/placement batteries and provide a more face valid pre-interview criterion.

FL016943

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. Schoffman

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Although indirect assessment of oral skills has been in use for some time (Lowe and Clifford 1980), the use of tape recorded question and answer formats in university-level oral skills assessment is now gaining popularity (Clark 1986). In Japan, where video hardware is readily available at virtually all institutions of higher learning, indirect assessment of oral skills has new potential in combination with voice recording equipment. For large-scale institutional testing, however, there are of course certain constraints which would apply to an assessment of a foreign language speaker's ability to narrate a short story utilizing the video as the medium of speech elicitation. Some of these are outlined below.

* Individual speech samples should be recorded on tape.
* Speech samples should be reliable.
* The speaking task should have multiple alternate forms.
* Language laboratory (L.L) assistants should be able to administer the test.
* Mass administration should be possible in a video-equipped L.L.
* Short random samples of the each recording should be scorable.
* Both holistic and atomistic scoring methods should be possible.
* Scores from the narrative discourse test should have high concurrent validity with face-to-face interviews.

On-Line Discourse

In the past decade only a few studies have explored phenomena associated with the production of narrative discourse (Chafe 1979, 1980; Ochs 1979). In a study involving the on-line narration of a short animated cartoon, Tomlin (1984) found that both native speakers and advanced non-native speakers were inclined to use a pragmatic mode of discourse which relied on left-dislocation and linear coordination of clauses, and in the case of non-native speakers of English, the non-use of grammatical morphology. Tomlin's study suggests that the use of a 'cold turkey' approach to narration, one in the unrehearsed extemporaneous mode, will cause a regression to a basic paratactic strategy by any speaker. In light of Tomlin's findings, such an approach to narration for the purposes of foreign language assessment would not be indicated since there would be an even more profound reduction of variance, especially if the sample of foreign language speakers to be tested is at the lower end of the ability scale.

In order to make oral narration a practical tool for foreign language speaking assessment, certain modifications to the extemporaneous production approach are necessary. Pre-viewing of the animated sequences in the learners' first language, for example, would allow each individual an equal starting point in terms of his or her anticipation of the outcome of the story line. The presentation of the complete story line, in the case of the pilot version of the project (an animated Japanese folk tale), allows each individual an equal opportunity to approach the speaking task in a culturally familiar framework.

One problem associated with the pre-viewing phase of the presentation of the narrative material arises, however, when individuals assume that they are to relate the subsequent scenes of the story from the perspective of each participating character. The resulting first-person narrative would be problematic in that the syntax manifest in its production is less elaborate than a third-person narration of the story line. This, then, necessitates a pre-viewing

presentation of an example of a third-person narration of an animated cartoon.

The next phase, by far the most challenging aspect of creating a video narrative task, is in selecting and editing scenes of equal length and salience so as to maximize the potential for differences in speaking skill to emerge. This editing phase requires an optimal balance between the need to present the narrative task so that production would be in 'real time' and at the same time allow the speaker to anticipate the up-coming sequence of events in the story line and plan her discourse accordingly.

Proficiency Scales

The primary use of the narrative discourse task in this study was for screening and placement into a five-level course in spoken English for college freshmen. Consequently, the anticipated ranges of proficiency levels were limited to known parameters within the institution: from persons with virtually no experience speaking a foreign language to persons with considerable experience in acquisition-rich environments. No need was seen, however, to establish a native speaker norm for the holistic proficiency scales. Instead, a small sub-set of learners with extensive overseas contact with English were chosen as the anchor population to which the scales could be developed in a descending order. The result of this approach to scaling was a set of anchor scores in three components of basic speaking proficiency: pronunciation, accuracy and fluency. The descriptors for each of the subscales follow:

Pronunciation	"4" Speaker can use reductions variably. Sentence level stress and accent are accurate.
	"3" Speaker primarily uses English syllable structures. Stress and accent are variable.
	"2" Speaker relies on Japanese syllable structure. English vowels are more prevalent than Japanese cardinal vowels.
	"1" Speaker uses Japanese syllable structure and cardinal vowels.
	"0" No response.
Accuracy	"5" Speaker uses a relatively wide range of syntactic patterns and uses accurate verb morphology.
	"4" Speaker can use a variety of syntactic patterns. Salient morphology is accurate.
	"3" Speaker can provide a canonical sentence pattern for each scene. Verb morphology is unstable.

"2" Speaker must string canonical sentences across more than one scene. Little verb morphology appears.

"1" Speaker can only manage to provide isolated words.

"0" No response.

Fluency

"6" Speaker can narrate the story at the same pace the scenes change, and can anticipate up-coming frames.

"5" Speaker can produce two or more T-units for some scenes. Multi-clause T-units appear.

"4" Speaker can manage a short T-unit for each scene.

"3" Speaker can complete T-units selectively. T-units may be started for each scene, but only longer scenes get completed T-units.

"2" Speaker can complete T-units by spreading them across two or more scenes.

"1" Speaker attempts utterances, but commonly abandons them when the scenes change.

"0" No response.

In addition to the above holistic ratings,[1] T-unit analyses were done on one hundred forty-eight of the recordings. Such a syntactic analysis of the recorded narrative was thought to be a viable method of determining the extent to which raters could differentiate between the accuracy and fluency criteria established in the holistic rating scales. An analysis which could provide empirically defined measures of accuracy and fluency would also give ample convergent and discriminant evidence for the holistic scales in that the atomistic T-unit analyses would serve to provide specific criteria with which comparisons of the holistic scales could be made. To this end T-unit analysis (see Gaies 1980, Larsen-Freeman 1978) was chosen because it is relatively transparent and does not require a great deal of rater training. However, the primary weakness in performing a T-unit analysis on samples of spoken narrative is the fact that the analysis of syntactic development begins at the clause level. Other approaches, such as that endorsed by Ferguson (1980), include sub-clause level criteria within tone groups as the primary units of analysis. A preliminary comparison of tone group analysis and T-unit analysis on the initial set of samples for rater

[1] All holistic ratings of pronunciation, fluency, and accuracy, as well as the atomistic T-Unit analyses, were done on recordings of the student narratives.

norming, however, suggested that tone groups corresponded to T-units at the clause level.[2] Given that the focus of the syntactic analysis was to establish operationalizable criteria for fluency and accuracy, no precision was considered lost in beginning with T-units as the basic unit of syntactic measurement. The three primary categories of T-units of interest in the analysis were total spoken T-units, total error-free T-units and words in error-free T-units.

Reliability

Two hundred and forty-seven recordings of the first of two parallel forms narrative tests were collected for the initial analysis. Random samples of ten narratives were collated on single cassettes and distributed to four raters for the primary inter-rater reliability estimation. Once a reasonable coefficient of reliability was established on the holistic rating (W = .85), the T-unit reliability phase could be undertaken. In all, three sets of ten samples each were rated before the holistic rating of the entire corpus of recordings was started. Owing to the time required to complete careful syntactic analyses of the narratives, a smaller sub-set of one hundred and forty-eight comprised the corpus for the more detailed analysis. The time spent in the initial rater norming phase turned out to be well spent. The Kendall's concordance coefficient for the four holistic raters reached .85, and the objective scoring required in the T-unit analysis by three raters was greater than .90 for counts of total T-units and greater than .95 for the error-free T-units.

Factor Analysis of Sub-Scores

Two versions of the narrative discourse task were prepared. The first, based on a Japanese folk tale about "the cunning of foxes", was presented to the incoming freshman class at the beginning of the academic year. The second version of the task, "the magic carp," was presented to the same students at the end of the nine-month academic term. Both versions were presented in a 'state of the art' language laboratory equipped with large video screens at the front of the laboratory and individual tape recording facilities in each cubicle. Since 247 students were to be tested for each version, six rotations in and out of the language laboratory had to be managed in a four-hour period. In both administrations, totaling four hundred and ninety-four recordings, only one tape was lost to a mechanical malfunction.

---

[2] In the framework introduced by Ferguson (1980), sub-clause level analysis of tone groups would assign weighted scores to well-formed verbs or nouns with functional words attached within the tone group. The T-Unit approach used in the present study was limited to dichotomous categorizations of T-Units (clauses) as +/-error-free. The tone group method would thus assign sub-scores for accuracy for a sentence like "I go/to the bank/this morning." In the T-Unit approach, however, such an utterance would be considered a single spoken T-Unit.

Immediately prior to the viewing of the animated cartoons, all students were given a short battery of paper-and-pencil tests. Listening cloze, dictation and a multiple choice grammar test were selected from a larger battery of tests developed in previous years. These tests were included for the purpose of comparison with the scores derived from the speaking tasks.

Two separate correlation matrices were generated from data derived from the two administrations of the paper-and-pencil tests and the narrative discourse tasks. Specifically, input to each matrix included scores from the speaking task: ratings for pronunciation, accuracy and fluency; objective measures of total T-units, error-free T-units, words within error-free T-units, the ratio of words in error-free T-units to total error-free T-units (WEFTU/EFTU), the ratio of words in error-free T-units to total T-units (WEFTU/TU), the ratio of error-free T-units to total T-units (EFTU/TU), and the ratio of words in error-free T-units to the number of error-free T-units times the total of T-units in the narrative (WEFTU/EFTU x TU). Listening cloze, dictation and multiple choice grammar scores represented the paper-and-pencil method of testing. Each matrix was entered into a separate principal factor analysis and rotated to a three factor solution. The purpose of the separate analyses was to establish the similarity between the two separate speaking tasks after the nine-month period of instruction.

Table 1

Factor Structure of Two Narrative Discourse Tasks

Factors	Eigenvalues		% of variance		Accumulative % var.	
	pre	post	pre	post	pre	post
1	7.03	7.25	66.6	67.2	66.6	67.2
2	1.71	1.50	16.2	13.9	82.8	81.1
3	.76	1.04	7.2	9.6	90.1	90.8
4	.50	0.66	4.7	6.1	94.8	97.0

Table 2

Orthogonal Rotation

	Loadings					
	factor 1		factor 2		factor 3	
	pre	post	pre	post	pre	post
pronunciation	.612	.496				
fluency	.855	.827				
accuracy	.672	.592		.577		
holistic total	.853	.744		.511		
dictation						.693
grammar					.446	.536
cloze					.445	.707
T-units	.887	.844				
error-free TU	.576	.492	.765	.855		
words in EFTU	.501	.417	.767	.855		
WEFTU/EFTU			.885	.941	.390	
WEFTU/TU		.413		.309	.665	.381
EFTU/TU			.939	.956		
WEFTU/EFTU x TU	.707	.745		.349		

Similar factor structures emerged on both versions of the narrative discourse test. Of main interest are the first two factors which account for about 82% of the variance on the two tests. Factor one consists of a single 'pure' measure of verbal fluency and complex measures that straddle the main fluency measure and the error-free measures. Factor two, whose most consistent measures are the ratio of words in error-free T-units to error-free T-units and the ratio of error-free T-units to total T-units, may be more a measure of morphosyntactic accuracy, given the fact that the measures containing the 'error-free' criterion load on this factor.

The holistic ratings partially support this dichotomy. The accuracy rating on the post-test loads with the objective accuracy measures as we might expect. One reason why the accuracy rating appears complex, with loadings on both factors on the post-test while only loading on the first factor on the pretest, may plausibly be traced to subtle changes in rating criteria by one or more of the raters. Once familiarized with the use of the objective measures of accuracy, i.e., those based on T-Unit analysis, the raters might have listened to the recordings counting error-free T-Units when they were rating the accuracy of the narratives on the post-test.

The fluency rating loads only on the first factor. The holistic total again loads with the first factor on the pretest and becomes more complex by the second administration of the narrative task, most likely because of the influence of the accuracy rating. The most striking feature of the holistic loadings, however, is the apparent inclination of the raters to recognize the gross verbosity produced by the speaker as a sign of a more general speaking proficiency. In contrast, there were a few students who remained relatively taciturn until they found a particular sequence in the narrative story in which

they could produce a perfectly formed utterance, and it appears likely that the raters initially recognized their silence as an overt sign of lack of proficiency.

The paper-and-pencil tests did not load on the two speaking factors. Given the fact that the multiple choice grammar test could not be empirically parsed out from the two measures of listening comprehension, dictation and the listening cloze, it is possible that these six tests were a reflection of a method of measurement artifact.[3]

Concurrent Validity

One of the goals of developing the narrative discourse test is to find a task that will share substantial variance with a more direct interview task. To this end several interview formats were considered in the concurrent validation phase of the project. The use of 'quick and dirty' interviews (i.e., interviews with a modicum of face validity and little reliability) was a temptation given the large numbers of potential interviewees and the unavailability of a sufficient number of native speaker interviewers. However, a structurally-based face-to-face interview was chosen to avoid the 'clam-up' phenomenon common to first interviews with Japanese students. It was believed that a structural describe-a-picture type of interview would bypass this problem. Subsequently, the John Test (Language Innovations n.d.) was chosen as the interview method because it contained sub-sections on structure and a freer narrative description of a short story depicted on the student's test material. The John Test was administered to a small non-random sample of students placed into a pre-intermediate level EFL course. It is important to note that since the interviewed group had been streamed, the likelihood of a dispersion of interview scores was precluded. Nonetheless a product moment correlation of .77 ($n = 44$) was found between the holistic rating and the John Test interview. The correlation would no doubt have been larger had students from the entire range of courses been interviewed.

A secondary goal in devising the narrative discourse task is to filter out less viable methods of rating samples of speech and find a single rating scheme that will provide the greatest amount of information about speaking proficiency in the least amount of time. An essential step in finding consistent measures is checking the pre and post-test correlations. The table below suggests that the measures of accuracy were the most erratic, presumably owing to the effects of instruction during the nine-month period between test administrations. The fluency measures, on the other hand, show relatively more stability over time.

---

[3] The dictations used in this study were short (five sentences long) and had lead-ins to each of the sentences. The cloze passages, while much longer, involved listening to a reading of each passage and allowed ample time for the filling in of the gaps. It appears possible that, despite the aural presentation of the dictation and cloze tests, most of these Japanese students relied on a gap-filling strategy first instead of attending to the aural cues. If indeed this strategy was widely used, the distinction between the listening and grammar tests may have been diminished.

Table 3

Correlations Between Pre and Post Test Measures

pronunciation	.284	T-units	.564	cloze	.484
fluency	.490	W/E x T	.595	dictation	.315
accuracy	.445	WEFTU/EFTU	.244	grammar	.489
holistic total	.509	WEFTU	.442	EFTU	.334

($p < .01$ for all correlations)

Conclusions

Given the impracticality of oral interviews in most foreign language testing contexts in Japan, screening for face-to-face interviews and placement tests can be facilitated with the use of narrative discourse tasks which can be recorded for analysis at a later date. In devising the tasks, care must be taken to provide appropriate background information to all of the examinees while creating speaking tasks that stimulate the speaker to narrate in 'real time'. The ultimate proof of the usefulness of the narrative discourse test, however, will be in its practicality; tape samples will by necessity be short and the rating criteria will need to be made more concise before this approach to testing oral proficiency can gain wider acceptance. The final phase of the experiment with the narrative discourse test will involve generalizability studies of progressively shortened samples of student speech so that the optimal intersection of reliability, validity and practicality can be estimated.

References

Chafe, W.L. 1979. The flow of thought and the flow of language. In T. Givon (Ed.), Discourse and syntax: syntax and semantics. 12, 159-182 New York: Academic Press.

Chafe, W.L. 1980. The pear stories: Cognitive, cultural and linguistic aspects of narrative production. Norwood, New Jersey: Ablex.

Ferguson, N. 1980. The Gordian Knot. Geneva, C.E.E.L.

Gaies, Stephen, J. 1980. T-unit analysis in second language research: applications, problems and limitations. TESOL Quarterly 14: 1.

Larsen-Freeman, Diane 1978. An ESL index of development. TESOL Quarterly 12: 4.

Ochs, E. 1979. Planned and unplanned discourse. In T. Givon (Ed.), Discourse and syntax: Syntax and semantics. 12, 51-80 New York: Academic Press.

Tomlin, R.S. 1984. The treatment of foreground-background information in the on-line descriptive discourse of second language learners. Studies in Second Language Acquisition. Vol 6 No. 2., 115-142.

The John Test (n.d.) Language Innovations Inc. 2112 Broadway Room 515. New York, N.Y.